1 Сравнение BigARTM с другими библиотеками

Сравниваем BigARTM с реализацией алгоритма Online VB LDA в библиотеке для тематического моделирования Gensim[1] и Vowpal Wabbit[2]. Для сравнения будем строить тематическую модель для документов из английской Википедии (корпус описан далее).

Перплексия. Алгоритм LDA представляет тематическую модель в виде распределений Дирихле на строчки Θ и столбцы Φ:

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\gamma}_d), \quad \boldsymbol{\phi}_t \sim \text{Dir}(\boldsymbol{\lambda}_t) \tag{1}$$

Для того чтобы сравнить перплексию матриц Φ, Θ для hold-out документов, мы будем брать средние распределения

$$\boldsymbol{\theta}_d^{\text{mean}} = \mathbb{E}_{\text{Dir}(\boldsymbol{\gamma}_d)}\boldsymbol{\theta}_d, \quad \boldsymbol{\phi}_t^{\text{mean}} = \mathbb{E}_{\text{Dir}(\boldsymbol{\gamma}_d)}\boldsymbol{\phi}_t. \tag{2}$$

Параметры эксперимента

- Машина: x86_64, 32 ядра, 1324.898MHz
- Corpus: English Wikipedia shapshot `2014-12-08`, hold-out 100'000 documents
- Topics: 100
- One pass through train documents of the corpus
- Batch size: 10'000 (`chunksize` in Gensim, `-minibatch` in VW)
- Update rule: $\rho = (\tau_0 + t)^{-\kappa}$, $\tau_0 = 1$, $\kappa = 0.5$
- Update after each batch in non-parallel implementation, update after P batches when running in P parallel threads (`update_every = num_processors`)
- LDA Priors: $\alpha = 0.1$, $\beta = 0.1$ ($\boldsymbol{\theta}_d \sim \text{Dir}(\alpha)$, $\boldsymbol{\phi}_t \sim \text{Dir}(\beta)$)

2 On LDA and ARTM

Данный раздел посвящён сравнению моделей LDA и ARTM. В экспериментах, сравнивающих BigARTM с Gensim и VW.LDA, мы показали, что алгоритм Online PLSA со сглаживающим регуляризатором и Online VB LDA работают схожим образом. Поэтому в этом эксперименте будут оцениваться характеристики PLSA со сглаживающим регуляризатором (который мы далее будем называть LDA) и ARTM (суть PLSA с набором регуляризаторов).

[1] `http://radimrehurek.com/gensim/`
[2] `https://github.com/JohnLangford/vowpal_wabbit/wiki/Latent-Dirichlet-Allocation`

Library	Proc.	Train Time	Inference Time	Perplexity
BigARTM Smoothing	1	62 min	127 sec	4000
Gensim LDA	1	369 min	395 sec	4161
Vowpal Wabbit LDA	1	73 min	120 sec	4108
BigARTM Smoothing	8	8 min	24 sec	4304
Gensim LDA-Multicore	8	70 min	338 sec	4470

Таблица 1: Сравнение BigARTM с реализацией LDA в библиотеке Gensim и Vowpal Wabbit. Train Time — время на обучение модели, Inference Time — время вычисления θ_d для всех документов из hold-out. Perplexity — перплексия посчитанная по обученной модели $\{\boldsymbol{\phi}_t\}$ на hold-out документах $\{vec\theta_d\}$, в случае Gensim и VW в качестве распределений $\boldsymbol{\phi}_t$ и $\boldsymbol{\theta}_d$ брались средние из соответствующих распределений Дирихле.

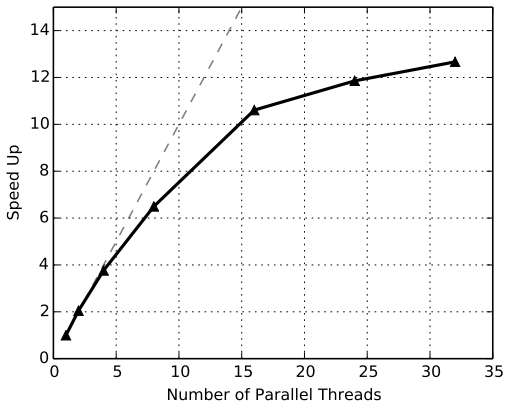

Рис. 1: BigARTM speed up

Текстовая коллекция Все наши эксперименты проводились на корпусе английской Википедии [3] , объём которой $|D| \approx 3.7 \times 10^6$ документов. Словарь имеет размер $|W| \approx 10^5$, общая длина коллекции в словах $n \approx 577 \times 10^6$.

Параметры эксперимента В этом эксперименте мы будем пользоваться следующими функционалами качества моделирования:

- Перплексия на контрольной выборке [4].
- Разреженность матрицы Φ.
- Разреженность матрицы Θ документов обучающей выборки.
- Характеристики ядер тем (размер, чистота, контрастность) ([**?**] ССЫЛКА НА НУЖНУЮ ПУБЛИКАЦИЮ КВ!!!).

Обе модели будут иметь следующий общий набор параметров, с которыми будет запускаться BigARTM: 1 проход по коллекции [5] , 10 проходов по каждому документу, 100 выделяемых тем. Матрица Θ, построенная на предыдущем проходе по документу, используется в качестве начального приближения на текущем. Параметры обновления матрицы Φ, κ и τ_0, равны 0.5 и 64 соответственно [6] . Порог $p(t|w)$ для ядровых функционалов — 0.25,. Размер батча равен 10000, обновления модели производится каждые батч.

Параметры LDA $\alpha = \beta = \dfrac{1}{|T|}$.

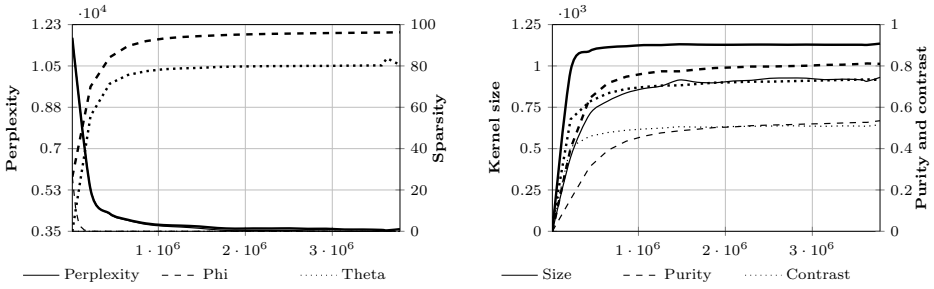

Рис. 2: Comparison of LDA (thin) and ARTM (bold) models. X axis is a number of processed documents.

Регуляризатор для ARTM, представляющий собой смесь разреживания и декорреляции тем, описывается формулой

[3] Коллекция была получена с помощью gensim.make_wikicorpus.

[4] Объём контрольной выборки, на которой перплексия измерялась в ходе прохода по коллекции — 10 тыс. документов. Кроме того, была измерена результирующая перплексия на выборке из 100 тыс. документов.

[5] Подразумевается один полный проход по всей коллекции и повторный проход по первым 1.5×10^5 документам для уточнения их распределений.

[6] Как это было в экспериментах в **??** РАЗДЕЛ ПРО СРАВНЕНИЕ БИБЛИОТЕК!!!

Таблица 2: Comparison of LDA and ARTM models. Quality functionals: $\mathcal{P}_{10k}$ $\mathcal{P}_{100k}$ — hold-out perplexity on 10.000 and 100.000 documents sets, $\mathcal{S}_\Phi$, $\mathcal{S}_\Theta$ — sparsity of Φ and Θ matrices (in %), $\mathcal{K}_s$, $\mathcal{K}_p$, $\mathcal{K}_c$ — average topic kernel size, purity and contrast respectively.

Model/Functional	$\mathcal{P}_{10k}$	$\mathcal{P}_{100k}$	$\mathcal{S}_\Phi$	$\mathcal{S}_\Theta$	$\mathcal{K}_s$	$\mathcal{K}_p$	$\mathcal{K}_c$
LDA	3499	3827	0.0	0.0	931	0.535	0.516
ARTM	3592	3944	96.3	80.5	1135	0.810	0.732

$$R(\Phi,\Theta) = -\beta \sum_{t\in T}\sum_{w\in W} \beta_w \ln \phi_{wt} - \alpha \sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} - \gamma \sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W} \phi_{wt}\phi_{ws} \to \max. \tag{3}$$

Отсюда получаются формулы М-шага

$$\phi_{wt} \propto \Big(n_{wt} - \beta \underbrace{\beta_w [t\in T]}_{\substack{\text{sparsing}\\ \text{topic}}} - \gamma \underbrace{[t\in T]\, \phi_{wt} \sum_{s\in T\setminus t} \phi_{ws}}_{\text{decorrelation}}\Big)_+; \tag{4}$$

$$\theta_{td} \propto \Big(n_{td} - \alpha \underbrace{\alpha_t [t\in T]}_{\substack{\text{sparsing}\\ \text{topic}}}\Big)_+. \tag{5}$$

Коэффициенты β_w и α_t примем равными 1, $\forall w, t$. Коэффициенты регуляризации α, β и γ возьмём постоянными на протяжении всего прохода по коллекции. Их значения: $\alpha = 0.15, \beta = 0.009, \gamma = 7.8 \times 10^5$.

Результаты В таблице 2 приведены финальные значения функционалов качества после одного прохода по коллекции для моделей LDA и ARTM. Видно, что комбинация регуляризаторов разреживания и декорреляции улучшает качество результирующей модели с небольшими потерями перплексии.

Более подробно процесс обучения представлен на 2. На верхнем графике показано убывание перплексии и замеры разреженностей матриц Φ и Θ. На нижнем — усреднённые характеристики ядер тем. Видно, что LDA совершенно не способствует разреживанию и даёт менее чистые и контрастные ядра тем, чем ARTM.